e-banking Prediction using Data Mining Methods

VASILIS AGGELIS
Department of Computer Engineering and Informatics
University of Patras
Rio, Patras
GREECE

PANAGIOTIS ANAGNOSTOU
Technological Education Institute of Patras
Patras
GREECE

Abstract: Data mining is a relatively new tendency in Informatics. Various companies and organizations adopt such methods in order to increase their profit and diminish their cost. The development and continuous training of prediction models is a very significant task, especially for bank organizations. The establishment of such models with the capacity of accurate prediction of future facts enhances the decision making and the fulfillment of the bank goals, especially in case these models are applied on specific bank units. E-banking can be considered such a unit receiving influence from a number of different sides. Application of data mining methods allows for successful prediction of e-banking parameters like the transactions volume conducted through this alternative channel in relation with other crucial parameters like the number of active users.

Key-words: Data Mining, Neural Networks, C&R Trees, Evaluation Charts, e-banking

1 Introduction

A number of data mining methods are nowadays widely used in treatment of financial and bank data [5]. The contribution of these methods in prediction making is essential. Study and prediction of bank parameter's behavior generally receives various influences. Some of these can be attributed to internal parameters while others to external financial ones. Therefore prediction making requires the correct parameter determination that will be used for predictions and finally for decision making.

Methods like Classification, Regression, Neural Networks and Decision Trees are applied in order to discover and test predictive models. In results validation, the contribution of bank specialists with knowledge and experience is essential for correct judgment of the produced results. The lack of advanced knowledge could lead in confusion and erroneous results.

Prediction making can be applied in specific bank data sets. An interesting data sector is e-banking. E-banking is a relatively new alternative payment and information channel offered by the banks. The familiarity of continuously more people to the internet use, along with the establishment of a feeling of increasing safety in conduction of electronic transactions imply that slowly but constantly e-banking users increase. Thus, the study of this domain is of great significance to a bank.

In the present study focus will be given in prediction making and testing concerning the transactions volume through e-banking in correlation to the number of active users. The software used is SPSS Clementine 7.0. A general description of prediction methods can be found in section 2. In section 3 the process of prediction making and testing is described while experimental results are discussed in section 4. Finally, in section 5 conclusions and future work are stated.

2 Classification, Regression and Neural Networks

There are two types of problems that can be solved by Classification and Regression methods [3].

Regression-type problems. The problem of regression consists of obtaining a functional model that relates the value of a response continuous variable Y with the values of variables $X_1, X_2, ..., X_v$ (the predictors). This model is usually obtained using samples of the unknown regression function.

Classification-type problems. Classification-type problems are generally those where one attempts to predict values of a categorical response variable from one or more continuous and/or categorical predictor variables.

2.1 Classification and Regression Trees (C&RT)

C&RT method [8, 12] builds classification and regression trees for predicting continuous dependent variables (regression) and categorical predictor variables (classification). There are numerous algorithms for predicting continuous variables or categorical variables from a set of continuous predictors and/or categorical factor effects. In most general terms, the purpose of the analysis via tree-building algorithms is to determine a set of *if-then* logical conditions that permit accurate prediction or classification of cases.

Tree classification techniques [1, 3, 15], when they "work" and produce accurate predictions or predicted classifications based on a few logical if-then conditions, have a number of advantages over many of those alternative techniques.

Simplicity of results. In most cases, the interpretation of results summarized in a tree is very simple. This simplicity is useful not only for purposes of rapid classification of new observations but can also often yield a much simpler "model" for explaining why observations are classified or predicted in a particular manner.

Tree methods are nonparametric and nonlinear. The final results of using tree methods for classification or regression can be summarized in a series of logical if-then conditions (tree nodes). Therefore, there is no implicit assumption that the underlying relationships between the predictor variables and the response variable are linear, follow some specific non-linear link function, or that they are even monotonic in nature. Thus, tree methods are particularly well suited for data mining tasks, where neither a priori knowledge is available nor any coherent set of theories or predictions regarding which variables are related and how. In those types of data analysis, tree methods can often reveal simple relationships between just a few variables that could have easily gone unnoticed using other analytic techniques.

2.2 Neural Networks

Neural networks [6, 8, 9, 16, 17] are non-linear predictive models that learn through training and resemble biological neural networks in structure. The advantage of this technique is that it can handle complex problems, with a large number of predictors that have many interactions. The results are not easy to interpret. This method requires that all variables are numeric. The neural network method is a good choice when the miner is more interested in the results of the model than in understanding how the model works.

Multilayer Perceptrons [14] is the most popular network architecture in use today. The units each perform a biased weighted sum of their inputs and pass this activation level through a transfer function to produce their output, and the units are arranged in a layered feedforward topology. The network thus has a simple interpretation as a form of input-output model, with the weights and thresholds (biases) the free parameters of the model. Such networks can model functions of almost arbitrary complexity, with the number of layers, and the number of units in each layer, determining the function complexity.

The number of input and output units is defined by the problem. The number of hidden units to use is far from clear. As good a starting point as any is to use one hidden layer, with the number of units equal to half the sum of the number of input and output units.

2.3 Evaluation Charts

Comparison and evaluation of predictive models in order for the best to be chosen is a task easily performed by the evaluation charts [4, 10, 11]. They reveal the performance of the models concerning particular outcomes. They are based on sorting the records as to the predicted value and confidence of the prediction, splitting the records into groups of equal size (quantiles), and finally plotting the value of the business criterion for each quantile, from maximum to minimum.

Different types of evaluation charts emphasize on a different evaluation criteria.

Gains Charts.

The proportion of total hits that occur in each quantile are defined as Gains. They are calculated as the result of: (number of hits in quantile/total number of hits) X 100%.

Lift Charts

Lift is used to compare the percentage of records occupied by hits with the overall percentage of hits in the training data. Its values are computed as (hits in quantile/records in quantile) / (total hits/total records).

Evaluation charts can also be expressed in cumulative form, meaning that each point equals the value for the corresponding quantile plus all higher quantiles. Cumulative charts usually express the overall performance of models in a more adequate way, whereas non-cumulative charts [4] often succeed in indicating particular problem areas for models.

The interpretation of an evaluation chart is a task certainly depended on the type of chart, although there are some characteristics common to all evaluation charts. Concerning cumulative charts, higher lines indicate more effective models, especially on the left side of the chart. In many cases, when comparing multiple models lines cross, so that one model stands higher in one part of the chart while another is elevated higher than the first in a different part of the chart. In this case, it is necessary to consider which portion of the sample is desirable (which defines a point on the *x* axis) when deciding which model is the appropriate.

Most of the non-cumulative charts will be very similar. For good models, noncumulative charts [4] should be high toward the left side of the chart and low toward the right side of the chart. (If a non-cumulative chart shows a sawtooth pattern, you can smooth it out by reducing the number of quantiles to plot and reexecuting the graph.) Dips on the left side of the chart or spikes on the right side can indicate areas where the model is predicting poorly. A flat line across the whole graph indicates a model that essentially provides no information.

Gains charts. Cumulative gains charts extend from 0% starting from the left to 100% at the right side. In the case of a good model, the gains chart rises steeply towards 100% and then levels off. A model that provides no information typically follows the diagonal from lower left to upper right (shown in the chart if Include baseline is selected).

Lift charts. Cumulative lift charts tend to start above 1.0 and gradually descend until they reach 1.0 heading from left to right. The right edge of the chart represents the entire data set, so the ratio of hits in cumulative quantiles to hits in data is 1.0. In case of a good model the lift chart, lift should start well above 1.0 on the left, remain on a high plateau moving to the right, trailing off sharply towards 1.0 on the right side of the chart. For a model that provides no information, the line would hover around 1.0 for the entire graph. (If Include baseline is selected, a horizontal line at 1.0 is shown in the chart for reference)

3 Generating predictions in e-banking data set

In the case of this study, as stated above, the objective is the production and test of predictions about the volume of e-banking transactions in relation to the active users. The term financial transactions stands for all payment orders or standing orders a user carries out, excluding transactions concerning information content like account balance, detailed account transactions or mini statement. The term «active» describes the user who currently makes use of the electronic services a bank offers. An active user may use e-banking services only for viewing deposits, mini statements and last transactions of his accounts. Active users are a subgroup of the enlisted users.

One day is defined as the time unit. The number of active users is the predictor variable (Count_Of_Active_Users) while the volume of financial transactions is assumed to be the response variable (Count_Of_Payments).

A sample of the above data set is shown in Table 1

Transaction Day	Count_Of_Active_Users	Count_Of_Payments
...	...	...
27/8/2002	99	228
28/8/2002	107	385
29/8/2002	181	915
30/8/2002	215	859
...	...	...

Table1 – Sample of Data

The date range for which the data set is applicable counts from April 20th, 2001 until December 12th, 2002. Data set includes data only for active days (holydays and weekends not included), which means 387 occurrences.

In order to create predictions Classification and Regression Tree (C&R Tree) [4, 13] was used as well as Neural Networks [4, 13] (prune method).

4 Experimental Results

4.1 C&RT

In Fig. 1 the conditions defining the partitioning of data discovered by the algorithm C&RT are displayed. These specific conditions compose a predictive model.

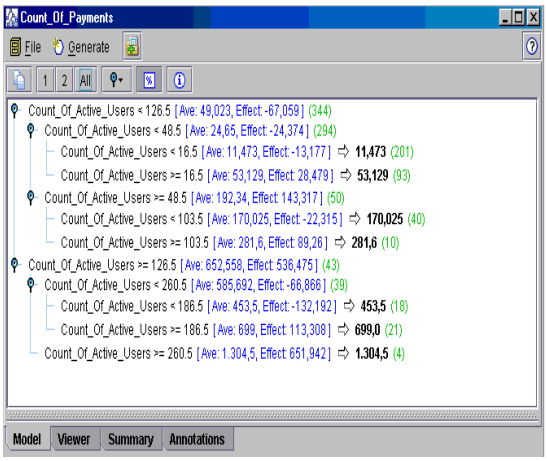

Fig. 1 - Conditions

C&RT algorithm [4] works by recursively partitioning the data based on input field values. The data partitions are called *branches*. The initial branch (sometimes called the *root*) encompasses all data records. The root is split into subsets or *child branches*, based on the value of a particular input field. Each child branch may be further split into sub-branches, which may in turn be split again, and so on. At the lowest level of the tree are branches that have no more splits. Such branches are known as *terminal branches,* or *leaves.*

Fig. 1 shows the input values that define each partition or branch and a summary of output field values for the records in that split.

For example a condition with many instances (201) is:

If Count_of_Active_Users < 16.5 then Count_of_Payments ⇨ 11,473,

meaning that if the number of active users during a working day is less than 16.5 then the number of transactions is predicted approximately at the value of 11.

Fig. 2 – Tree Structure

Fig. 2 shows a graphical display of the structure of the tree in detail.

Goodness of fit of the discovered model was assessed by the use of evaluation charts. Examples of evaluation Charts (Fig. 3 to Fig. 4) are shown below.

Gains Chart

This Chart (Fig. 3) shows that the gain rises steeply towards 100% and then levels off. Using the prediction of the model, the percentage of Predicted Count of Payments for the percentile is calculated and these points create the lift curve. An important notice is that the efficiency of a model is increased with the area between the lift curve and the baseline.

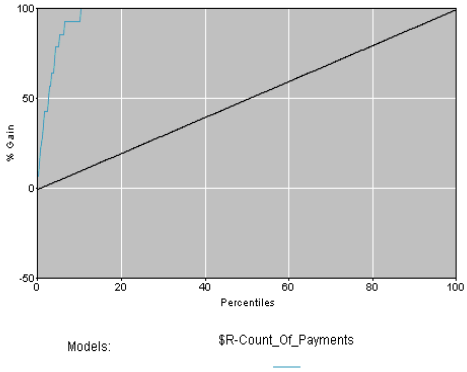

Fig. 3 – Gains Chart

Lift Chart

As can be seen in Fig. 4, Chart starts well above 1.0 on the left, remains on a high plateau moving to the right, and then trails off sharply towards 1.0 on the

right side of the chart. Using the prediction of the model shows the actual lift.

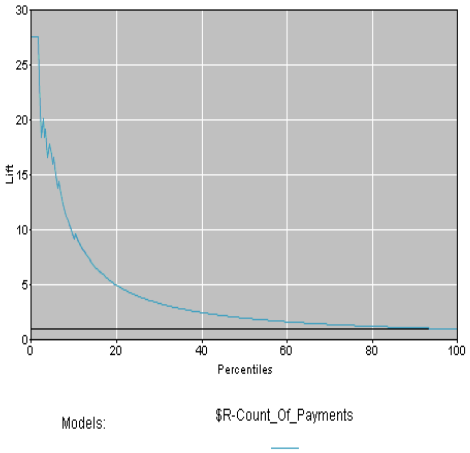

Fig. 4 – Lift Chart

In Fig. 5 the Count of Payment and the Predicted Count of Payment as functions of the number of active users can be seen. The Predicted Count of Payment line clearly reveals the partitioning of the data due to the application of the C&RT algorithm.

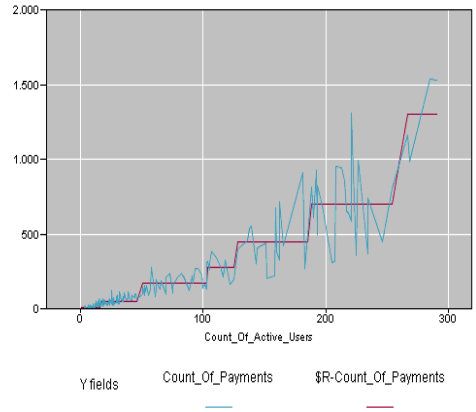

Fig. 5 - Real vs Predicted

4.2 Neural Networks

In order to certify the correctness and accuracy of the model, the method of Neural Networks was applied to the data set. The estimated prediction accuracy is 97.51%.

For numeric targets, the calculation is based on the differences between the predicted values and the actual values in the training data. The formula for calculation of the accuracy for numeric fields is:

(0.5 – (abs(Actual – Predicted)/(Range of Output Field))) * 100%

where *Actual* is the actual value of the output field, *Predicted* is the value predicted by the network, and *Range of Output Field* is the range of values for the output field (the highest value for the field minus the lowest value) [4]. Accuracy is calculated for each record, and the overall accuracy is the average of the values for all records in the training data set.

Because these estimates are based on the training data, they are likely to be somewhat optimistic. The accuracy of the model on new data will usually be somewhat lower than this.

The Relative Importance of Count of Active Users is 0,83241, implying that the number of active users is very important in the model's formation. The value listed for each input is a measure of its relative importance, varying between 0 (a field that has no effect on the prediction) and 1.0 (a field that completely determines the prediction).

In Fig. 6 the graph of the Count of Payment and the Predicted Count of Payment in this case as functions of the active users is presented. As can be observed the line of the Predicted Count of Payment is much smoother than the actual one.

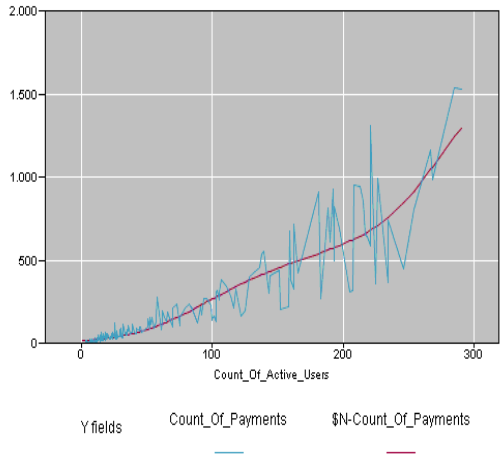

Fig. 6 – Real vs Predicted

Finally, the Linear Correlation of the model is 0.933. Since this value approaches unity it indicates a strong positive relation, such that high predicted values are associated with high actual values and vice versa.

5 Conclusions and Future Work

In this study the establishment of a prediction model concerning the number of payments conducted through Internet as related to the number of active users is investigated along with testing its accuracy.

Using the methods C&RT and NN it was concluded that there exists strong correlation

between the number of active users and the number of payments these users conduct. It seems that the above conclusion can be logically derived, but as we mentioned before it is not standard that an active user conducts payments. It is clear that the increase of the users' number results in increase of the transactions made. Therefore, the enlargement of the group of active users should be a strategic target for a bank since it increases the transactions and decreases its service cost. It has been proposed by certain researches that the service cost of a transaction reduces from €1.17 to just €0.13 in case it is conducted in electronic form. Therefore, a goal of the e-banking sector of a bank should be the increase of the proportion of active users compared to the whole number of enlisted customers from which a large number do not use e-services and is considered inactive.

Future work includes the creation of prediction models concerning other e-banking parameters like the transactions value, the use of specific payment types, commissions of electronic transactions and e-banking income in relation to internal bank issues. Also extremely interesting is the case of predictive models based on external financial parameters.

The prediction model of this study could be determined using also other methods of data mining. Use of different methods offers the ability to compare between them and select the more suitable. The acceptance and continuous training of the model using the appropriate Data Mining method results in extraction of powerful conclusions and results.

References:

[1]. L. Breiman, J. H. Friedman, R.A. Olshen, and C.J. Stone. "*Classification and Regression Trees*", Wadsworth Publications, 1984

[2]. A.Dobra. "*Classification and Regression Tree Construction*", Thesis proposal, Departmant of Computer Science, Cornell University, Ithaca NY, 2002

[3]. D. Hand, H. Mannila, and P. Smyth. "*Principles of Data Mining*". The MIT Press, 2001.

[4]. "*Clementine 7.0 Users's Guide*". Integral solutions Limited, 2002.

[5]. D. Foster, and R. Stine. "*Variable Selection in Data Mining: Building a Predictive Model for Bankruptcy*", Center for Financial Institutions Working Papers from Wharton School Center for Financial Institutions, University of Pennsylvania, 2002.

[6]. P. Cerny. "*Data mining and Neural Networks from a Commercial Perspective*", ORSNZ Conference Twenty Naught One, 2001

[7]. A. Comrie. "*Comparing Neural Networks and Regression Models for Ozone Forecasting*", The Journal of the Air & Waste Management Association, 1997

[8]. J. Galindo. "*Credit Risk Assessment using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications*", Computational Economics Journal, 1997

[9]. T. Hellstrom, and K. Holmstrom "*Predicting the Stock Market*", Techical Report Series, Center of Mathematical Model, Sweden, 1998

[10]. S.J. Hong, and S. Weiss. "*Advances in Predictive Model Generation for Data Mining*" Proceedings 1st International Workshop Machine Learning and Data Mining in Pattern Recognition, 1999

[11]. B. Zupan, J. Demsar, M. Kattan, M. Ohori, M. Graefen, M. Bohanec, and J.R. Beck. "*Orange and Decisions-at-Hand: Bridging Predictive Data Mining and Decision Support*" Workshop Integrating Aspects of Data Mining, Decision Support and Meta-Learning, 2001

[12]. R.Lewis. "*An Introduction to Classification and Regression tree (CART) Analysis*", Annual Meeting of the Society for Academic Emergency Medicine in San, 2000.

[13]. S.A. Madeira. "*Comparison of Target Selection Methods in Direct Marketing*", MSc Thesis, Technical University of Lisbon, 2002

[14]. C. Lu, J. De Brabanter, S. Van Huffel, I. Vergote, and D. Timmerman. "*Using Artificial Neural Networks to Predict Malignancy of Ovarian Tumors*", 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2001

[15]. A. Buja, and Y. Lee. "*Data Mining Criteria for Tree-based Regression and Classification*", The Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2001

[16]. M. Kon, and L. Plaskota. "*Complexity of Predictive Neural networks*", International Conference on Complex Systems, 2000.

[17]. M. Craven, and J. Shavlik. "*Using Neural Networks for Data Mining*", Future Generation Computer Systems Journal, 1997